Nikola Jerković [56]
Technical College of Applied Sciences in Zrenjanin, Zrenjanin/Serbia
Tanja Sekulić [57]
Technical College of Applied Sciences in Zrenjanin, Zrenjanin/Serbia
Ivana Berković [58]
University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin/Serbia

APPLICATION OF LARGE LANGUAGE MODELS IN DECISION-MAKING THEORY: ACCURACY, INTERPRETABILITY, AND LIMITATIONS

DOI: 10.5937/PIM26201J

Abstract:
The development of large language models (LLMs), such as ChatGPT, has opened new possibilities for supporting decision-making across various domains. This paper examines the potential application of ChatGPT version 5.3 in solving problems from the field of decision-making theory, with a focus on accuracy, interpretability, and limitations. Through case studies and the analysis of common problem types, the paper shows that LLMs can be a significant aid in solving formal tasks, but also exhibit systematic errors and biases. It is concluded that ChatGPT represents a useful support tool, but not a substitute for expert decision-making.

Keywords: *LLM, ChatGPT 5.3, decision-making theory*

Introduction

In the past decade, we have witnessed the rapid development of large language models (LLMs), which have significantly enhanced the capabilities of natural language processing and interaction between humans and computer systems. Such models, based on advanced deep learning techniques and trained on large datasets, have demonstrated outstanding performance in tasks such as text generation, translation, and answering questions [1].

With the evident advancement of these technologies, interest in their application to more complex cognitive tasks has also increased, among which decision-making holds a prominent place. Decision-making theory represents an important field that encompasses formal methods and models for selecting the optimal alternative under conditions of certainty, risk, and uncertainty [2]. In this context, LLMs

[56] nikolajerkovic13@gmail.com, ORCiD: 0009-0005-0978-6016
[57] tsekulicvts@gmail.com, ORCiD: 0000-0002-0977-4964
[58] ivana.berkovic62@gmail.com, ORCiD: 0000-0003-2366-3942

are increasingly being considered as potential tools for supporting problem analysis, data interpretation, and the generation of possible solutions.

One of the most well-known representatives of this class of models is ChatGPT, whose newer versions demonstrate significant progress in understanding context and generating coherent and structured responses [3, 4]. However, despite their impressive capabilities, the question remains to what extent these models can be reliably used in formal and mathematically grounded disciplines such as decision-making theory.

The main objective of this paper is to examine the applicability of the ChatGPT 5.3 model in solving decision-making theory problems, with particular emphasis on three key aspects: the accuracy of the obtained solutions, the interpretability of the generated explanations, and the limitations of the model when dealing with this type of problem. Through the analysis of specific examples and different approaches to task formulation, the paper aims to provide a realistic assessment of the capabilities and shortcomings of this tool.

1. Theoretical Background

The modern era, marked by the previously mentioned technological advancements, has led to the gradual overcoming of the strict application of traditional approaches based on mathematical models. Instead of the exclusive use of classical mathematical procedures, large language models (LLMs) have emerged as a complement in problem analysis and the formulation of potential solutions, introducing a new dimension due to their ability to interpret textually formulated problems.

Previous research has shown that large language models can successfully solve certain types of analytical and logical tasks, but with significant limitations. For example, studies indicate that models often exhibit a tendency to generate plausible yet incorrect answers, which is described in the literature as the phenomenon of so-called “hallucinations” [5]. In this regard, it is also important to consider the interpretability of the obtained responses, as the generated explanations do not necessarily reflect a true understanding of the problem, but may instead arise from statistical patterns in the data [6]. Furthermore, it has been established that the outputs of the models significantly depend on prompt formulation [7], where even simple

changes in the phrasing of the input text can lead to substantially different results [8].

In order to mitigate these limitations and improve reliability in more complex analytical processes, techniques such as "chain-of-thought" reasoning have been developed, which involve the explicit decomposition of a problem into a sequence of intermediate steps in the reasoning process. Such approaches have demonstrated significant improvements in solving mathematical and logical problems [9]. However, despite this, models can still make errors in multi-step calculations and exhibit inconsistency in their results, which represents a significant challenge for their application in domains such as decision-making theory.

2. Research Description

In this study, the application of large language models in solving problems from the field of decision-making theory was analyzed, with a focus on the ChatGPT 5.3 model. The selection of this model was based on its wide availability and advanced natural language processing capabilities, which make it suitable for the analysis of formal problems and the generation of solutions.

The experimental part of the study includes solving typical tasks addressed within decision-making theory. Three groups of problems were considered:

- decision-making without a priori probabilities;
- decision-making with a priori probabilities;
- decision-making under risk using Bayesian probabilities.

Particular attention was given to the way problems were presented to the model through textual instructions (prompts). In different iterations, various approaches to task formulation were used as needed, including:

- basic problem formulation without additional explanations;
- requests for step-by-step solutions with explanations;
- providing specific formulas;
- supplying a solved example from the literature as a guide;
- requests to solve only a single part of the task.

The aim of this approach was to examine the extent to which prompt formulation influences the generation of responses.

The evaluation of results was based on three key criteria:

- accuracy – by comparing the obtained results with correct solutions from the literature and teaching practice;
- interpretability – by assessing the clarity, logical coherence, and consistency of the explanations provided by the model;
- limitations – by identifying systematic and random errors, as well as inconsistencies across different solution iterations.

3. Research Results and Discussion

The research results indicate that ChatGPT 5.3 can successfully solve certain types of decision-making theory problems, but with significant limitations that depend on the complexity of the problem and the manner in which it is presented.

In problems involving decision-making without a priori probabilities, the model demonstrated a high level of accuracy. The calculations were largely correct, with one observed error related to the application of the Laplace criterion, where certain values were incorrectly computed. In the observed case, this error did not affect the final selection of the optimal alternative. It is important to emphasize that the model was not provided with additional instructions for solving the task, which suggests that, in this case, it can serve as a useful tool for solving simpler problems.

In the case of decision-making with a priori probabilities, the results were completely accurate. The model correctly applied expected values and made the correct decision, without any observed errors in calculation or reasoning.

The most significant issues were observed in tasks involving conditional (Bayesian) probabilities. In the first iteration, the model did not successfully solve the task. In subsequent attempts, despite additional instructions, inconsistencies appeared in the calculations, errors in the payoff tables, and an inability to properly use sample information for decision-making.

Interestingly, increasing the amount of instructions led to a higher number of errors, indicating limitations of the model in handling more

complex tasks. Even when explicitly provided with the formula for Bayesian probability, the model repeated the same incorrect results.

In later iterations, when a solved example from the literature was provided, the model partially improved its performance – it correctly calculated the posterior probabilities, but still made errors when selecting the optimal action. In the final iteration, further deterioration occurred, where even the basic calculations were no longer correct.

A comparative analysis of the obtained results across different problem types and evaluation criteria is presented in Table 1.

Table 1: *Comparative analysis of ChatGPT 5.3 performance by problem type and evaluation criteria*

Problem type / Evaluation criterion	***Accuracy***	***Interpretability***	***Limitations***
Decision-making without a priori probabilities	High, with minor calculation errors	Clear and consistent explanations	Incomplete computational accuracy
Decision-making with a priori probabilities	Complete, without calculation errors	High clarity and logical coherence	No significant limitations observed
Decision-making under risk using Bayesian probabilities	Low, with frequent errors across iterations	Explanations are inconsistent across iterations and may appear valid, although the obtained solutions are incorrect	Errors in calculations, high sensitivity to prompt formulation, and inconsistency across repeated runs

Source: Authors

4. Conclusion and Future Work

Based on the conducted research, it can be concluded that ChatGPT 5.3 represents a useful tool for supporting the solution of problems in the field of decision theory, but with clearly expressed limitations.

With regard to the criterion of accuracy, the model demonstrates a high level of reliability in simpler tasks and problems with clearly defined steps. However, in more complex problems that require multiple stages of solution development, accuracy significantly declines.

In terms of interpretability, the model often provides understandable and logically structured explanations, but these are not always consistent with the correct solutions. This means that the user may gain the impression of correctness even when the response contains errors.

The analysis of limitations shows that the model makes basic computational errors, has difficulties with complex multi-step problems, exhibits inconsistency across different iterations, and depends on the formulation of the prompt.

Based on these findings, it can be concluded that ChatGPT 5.3 can be useful as an auxiliary tool in education and preliminary problem analysis, but it cannot replace a human expert in decision-making, especially in situations that require high precision and reliability.

The limitations of this research are reflected in the relatively small number of analyzed tasks and prompts, as well as in the focus on a single model.

Future research may include:

- a broader set of problems from decision-making theory;
- a systematic analysis of the impact of different types of prompts;
- a comparison of multiple LLM models;
- a quantitative evaluation of performance across a larger number of experimental cases.

Further development of such tools and their proper integration into decision-making processes can significantly improve the efficiency of analysis, but only with the necessary retention of human oversight and expertise.

REFERENCES

[1] Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877-1901. doi:10.48550/arXiv.2005.14165

[2] Parmigiani, G., & Inoue, L. (2009). *Decision theory: Principles and approaches*. John Wiley & Sons.

[3] Qureshi, R., Shaughnessy, D., Gill, K. A., Robinson, K. A., Li, T., & Agai, E. (2023). Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation?. *Systematic Reviews*, *12*(1), 72. doi:10.1186/s13643-023-02243-z

[4] Makridakis, S., Petropoulos, F., & Kang, Y. (2023). Large language models: Their success and impact. *Forecasting*, *5*(3), 536-549. doi:10.3390/forecast5030030

[5] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1-38. doi:10.1145/3571730

[6] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. *In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). doi:10.1145/3442188.3445922

[7] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1-35. doi:10.1145/3560815

[8] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, *35*, 22199-22213. doi:10.52202/068431-1613

[9] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824-24837. doi:10.52202/068431-1800